SA-SDR: A novel loss function for separation of meeting style data

Thilo von Neumann, Keisuke Kinoshita, Christoph Boeddeker, Marc Delcroix and Reinhold Haeb-Umbach

Paderborn University, Germany NTT Corporation, Japan

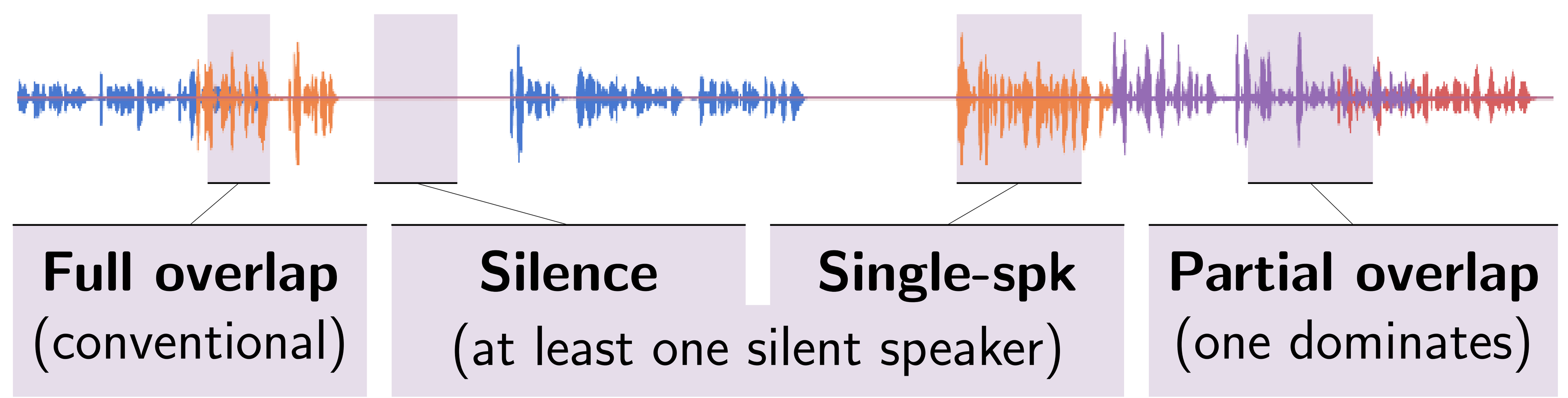
Introduction

The *averaged* Signal-to-Distortion Ratio (A-SDR) is a widely used objective function *(maximized)* for source separation

- **Problem**: A-SDR is not optimal for meeting scenarios
- **Goal**: Make the SDR more robust for meeting-like data

Meeting style data

Meeting style data is more challenging than conventional fully overlapped mixtures:



- Many active speakers
- Varying speaking patterns

Conventional: Averaged SDR (A-SDR)

Conventional objective used in many works, e.g., TasNet:

$$\text{A-SDR} = \frac{10}{K}\sum_{k=1}^{K}\log_{10}\frac{\|\mathbf{s}_k\|^2}{\|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2}$$

$\mathbf{s}$: Reference signal, $\hat{\mathbf{s}}$: Estimated signal, k: speaker index

(At least one) silent reference signal

$$\text{A-SDR} = \frac{10}{K}\sum_{k=1}^{K}\log_{10}\frac{\mathbf{0}}{\|\mathbf{0} - \hat{\mathbf{s}}_k\|^2} \rightarrow -\infty$$

- *undefined!*

Partial overlap / One dominating speaker

- The term of the *already well separated output* dominates

$$\text{A-SDR} \propto \underbrace{\log_{10}\frac{\left\|\mathbf{s}^{(\text{good})}\right\|^2}{\left\|\mathbf{s}^{(\text{good})} - \hat{\mathbf{s}}^{(\text{good})}\right\|^2}}_{\text{dominates } (\rightarrow\infty)} + \underbrace{\log_{10}\frac{\left\|\mathbf{s}^{(\text{bad})}\right\|^2}{\left\|\mathbf{s}^{(\text{bad})} - \hat{\mathbf{s}}^{(\text{bad})}\right\|^2}}_{\text{gets overruled}}$$

- Also the gradiets focus on the *already well separated output*

$$|\nabla_{\hat{\mathbf{s}}^{(\text{good})}}\text{A-SDR}| > |\nabla_{\hat{\mathbf{s}}^{(\text{bad})}}\text{A-SDR}|$$

Common 'hacks'

Many distort the loss value and/or heavily depend on hyperparameters

Here shown for a single pair of reference $\mathbf{s}$ and estimation $\hat{\mathbf{s}}$

Soft Maximum (*ε-thresholded SDR)*

$$\varepsilon\text{-tSDR} = 10\log_{10}\frac{\|\mathbf{s}\|^2 + \varepsilon}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + \tau(\|\mathbf{s}\|^2 + \varepsilon)}$$

- Prevents (to some degree) the overruling issue

Skewed SDR

$$\text{skewed SDR} = 10\log_{10}\frac{\|\mathbf{s}\|^2}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + \nu\|\hat{\mathbf{s}}\|^2}$$

log-MSE

$$\text{log1p-MSE} = -\log_{10}(\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + 1)$$

- Works for silent targets

Switch objective function for silent references

$$\mathcal{L}_0 = -10\log_{10}(\|\hat{\mathbf{s}}\|^2 + \tau\|\mathbf{y}\|^2)$$

- Loss has to be switched when a target is silent

Source-Aggregated SDR (SA-SDR)

Aggregate energies at source level instead of losses:

$$\text{SA-SDR} = 10\log_{10}\frac{\sum_{k=1}^{K}\|\mathbf{s}_k\|^2}{\sum_{k=1}^{K}\|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2}.$$

$\mathbf{s}$: Reference signal, $\hat{\mathbf{s}}$: Estimated signal, k: speaker index

Silent reference signals

$$\text{SA-SDR} = 10\log_{10}\frac{\|\mathbf{s}_1\|^2 + \mathbf{0}}{\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|^2 + \|\mathbf{0} - \hat{\mathbf{s}}_2\|^2}$$

- Stalbe when at least one reference signal is not silent

Partial Overlap / One dominating speaker

- The distortions of the well separated output disappear

$$\text{SA-SDR} \propto \log_{10}\frac{\left\|\mathbf{s}^{(\text{good})}\right\|^2 + \left\|\mathbf{s}^{(\text{bad})}\right\|^2}{\underbrace{\left\|\mathbf{s}^{(\text{good})} - \hat{\mathbf{s}}^{(\text{good})}\right\|^2}_{\text{disappears } (\rightarrow 0)} + \left\|\mathbf{s}^{(\text{bad})} - \hat{\mathbf{s}}^{(\text{bad})}\right\|^2}$$

- The gradients focus on the *not-so-well-separated output(s)*

$$|\nabla_{\hat{\mathbf{s}}^{(\text{good})}}\text{SA-SDR}| < |\nabla_{\hat{\mathbf{s}}^{(\text{bad})}}\text{SA-SDR}|$$

Experiments: WSJ0-mix

Loss	BSSEval SDR	A-SDR	SA-SDR
no separation	0.2	0.0	0.0
A-SDR	17.8	17.5	17.8
A-tSDR	17.8	17.5	17.8
SA-SDR	**18.0**	**17.7**	**18.0**
SA-tSDR	17.7	17.5	17.8

- A-SDR and SA-SDR have a comparable performance on fully overlapped data

Experiments: Meeting style data

Loss	#spk train	Metrics				
		WER	atten. ratio	BSSEval SDR	VAER	SA-SDR
no separation	—	48.1	0.0	7.3	65.6	0.0
A-SDR	2	13.5	25.5	19.1	12.6	13.8
A-log-MSE	2	13.1	18.3	19.5	13.2	14.8
A-log1p-MSE	1+2	13.5	25.3	19.6	9.9	16.8
A-skewed-SDR	2	15.6	24.7	18.7	12.5	10.1
A-tSDR	2	13.6	21.1	18.8	14.0	13.3
A-ε-tSDR	1+2	12.8	25.9	19.6	11.8	15.5
A-log-tMSE+$\mathcal{L}_0$	1+2	12.8	26.4	19.6	10.7	14.5
SA-SDR	1+2	12.5	30.3	19.8	9.7	16.1
SA-log-MSE	1+2	13.3	**31.5**	19.3	11.6	14.7
SA-log1p-MSE	1+2	15.1	25.1	18.7	11.4	15.7
SA-skewed-SDR	1+2	15.1	28.9	18.6	12.6	10.6
SA-tSDR	1+2	**12.2**	30.8	**19.9**	**8.2**	**17.9**
SA-ε-tSDR	1+2	12.8	27.5	19.6	9.1	16.3

- SA-SDR can reconstruct silence better than A-SDR (improvement in attenuation ratio, VAER and SA-SDR)
- Separation in overlapping regions is often comparable and sometimes better (similar WER)

Conclusions

- Stabilizing the loss often improves performance
- SA-SDR elegantly stabilizes the SDR for meeting style data without hyperparameters

Computational resources were provided by the Paderborn Center for Parallel Computing.